图书情报工作
第62卷 第1期 2018年1月

基于事件要素的组合模型微博热点事件摘要提取*

■ 李纲 徐伟 王馨平
武汉大学信息管理学院 武汉 430072

摘要：［**目的/意义**］为帮助读者从热点事件产生的海量微博报道中快速了解事件的来龙去脉，提高微博事件摘要的准确性和可读性，提出一种基于事件要素的多模型微博热点事件时间轴摘要提取方法。［**方法/过程**］针对微博文本特征，结合主题模型（LDA）与互信息最大熵模型（MaRxEnt-MI）的特点提取事件摘要关键词，以微博传播价值和主题相关性为标准筛选微博，以时间-摘要关键词-摘要微博的形式生成时间轴摘要。［**结果/结论**］利用人工标注的测试集，与传统的TextRank方法进行对比，F值提高8%-13%，内部测试表明摘要可读性提高明显。实验文本和测试集的数量及事件丰富度需要进一步扩展，应考虑更多的加权策略模型以提高摘要的准确性。实验结果及测试反馈表明，本文的方法能很好满足用户对热点事件摘要信息需求，提高微博摘要提取的准确率。

关键词：文本挖掘 事件摘要 潜在狄利克·雷分布 互信息最大熵模型
分类号：TP391
DOI：10.13266/j.issn.0252-3116.2018.01.013

1 引言

随着互联网的普及和微博的流行，普通网民、网络名人、新闻媒体以及政府机构等各种用户都将微博作为获取新闻信息、发表评论的主要途径。热点事件发生后，在微博平台上积累了海量的数据。然而由于发布的信息口语化严重、文本长度较短、语义缺失严重、垃圾文本多、信息增长量过快等特点，读者难以快速了解事件的来龙去脉。对于热点事件，时间要素是对事件描述的重要一环，在重要时间节点抽取能够表示热点事件发展情况的文本，可使用户通过这些文本能快速了解热点事件。

国内外关于自动文档摘要以及新闻摘要等技术的研究为解决上述问题提供了很多参考。自动文档摘要技术最早是应用于科技论文领域，是由于科技领域论文格式严谨、用语规范、数据的结构非常完整，论文本身的摘要也便于实验结果的验证[1]。2000年后，学界对自动摘要更加关注，如Document Understanding Conference（DUC）等的学术会议推动了研究的深入。传统的文档摘要方法按照实现的技术可以分为抽取式的摘要和生成式的摘要。然而，基于语义理解的方法虽然可读性好，但是语义语法分析实现起来复杂，应用领域的可移植性差、对文本质量要求很高。对于微博这样数据稀疏、表达不严谨、口语化严重的语料处理起来效率很低，实现难度也大。另一方面，基于统计文本特征抽取句子作为摘要的方法在技术上虽然容易实现、而且应用领域广泛，但是大多数研究忽略了时间的相关性，导致摘要不简洁、内容不全面、表达不连贯。总地来说，抽取式的方法比生成式的方法更适合微博事件的摘要生成[2-5]。

近年来，微博文本自动摘要技术逐渐兴起，国外相关研究如D. Inouye提出通过组合的TF-IDF算法对句子评分、排序去除冗余生成多条微博摘要，以及先通过微博聚类分析、然后抽取每个类别重要的微博作为摘要的方法[6]。R. Swan和J. Allan通过人工设计事件表，针对每个时间节点抽取命名实体，并以时间阶段串联起来作为事件年表[7]。国内研究则起步较晚，R. Long等基于关键词图聚类的方法选择微博热点事件内容相关的n条微博作为摘要[8]。X. Wan考虑到时间因素在文本结构中的影响，提出结合时间要素的

* 本文系国家社会科学基金重大项目"面向学科领域的网络信息资源深度聚合与服务研究"（项目编号：12&ZD221）研究成果之一。
作者简介：李纲（ORCID：0000-0001-5573-6400），教授，博士生导师；徐伟（ORCID：0000-0003-1060-293X），硕士研究生，通讯作者，E-mail：845310328@qq.com；王馨平（ORCID：0000-0003-3763-5483），硕士研究生。
收稿日期：2017-05-10 修回日期：2017-10-09 本文起止页码：96-105 本文责任编辑：王传清

TimeTextrank 文本摘要算法[9]。相关的研究大多数以 Twitter 作为研究对象,而且很少考虑时间因素的重要性[10－11]。新闻媒体摘要研究最先重视时间要素,但是微博事件摘要中相关的研究较少。

笔者通过分析微博上热点事件以及相关新闻报道,发现相关新闻报道中事件主体、事件中的时间、地点、人物是人们最关注的事件要素。从网民评论和关注点来看,用户询问的最多是事件发展过程中重要时间节点上的发展情况。从相关微博的内容上看,各重要时间段内微博中的事件要素、实体词的分布也是不断变化的。而且整体数据稀疏问题比较大、文本数据量大、口语化严重、无用信息多。

因此,本文提出基于 LDA(latent dirichlet allocation)模型提取主题关键词集[12],通过最大熵互信息模型解决主题模型提取关键词的无序性的缺点对主题关键词表进行优化,融合事件要素和微博影响力的综合衡量方法判断重要时间点内微博的重要性,从而得出热点事件的时间轴摘要。

2 研究思路与相关模型

2.1 研究思路

热点事件时间轴摘要是在事件几个重要的时间段上输出可以代表该节点事件发展情况的文本集,这些文本可以较全面地概括该时间节点网络上针对该事件新闻报道的主要内容。时间轴摘要过程包括重要事件特征抽取,时间摘要关键词抽取、摘要句子输出 3 个主要部分。

本文首先利用 LDA 模型提取微博语料的热点话题以及话题下的关键词,然后结合事件要素抽取结果、词性、在话题下的概率、MaxEnt-MI 计算词语间关系度等要素,计算关键词权重,生成摘要关键词。然后针对每个时间段内的微博,根据前面生成的摘要关键词结合微博本身的新闻价值计算权重,选取权重高的句子作为摘要生成,按照时间－摘要关键词－摘要微博的摘要形式呈现。具体流程见图 1。

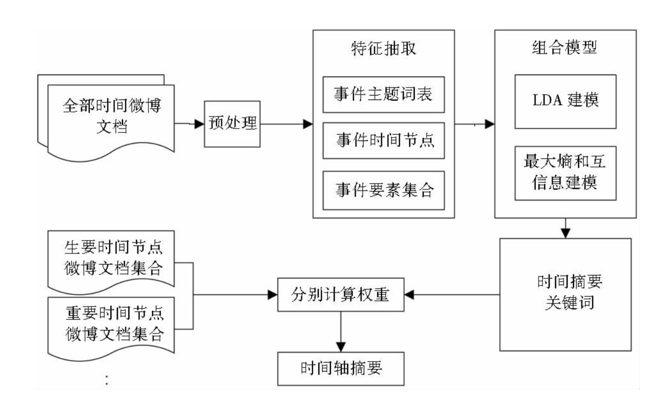


图 1 热点事件摘要生成流程

2.2 微博热点事件摘要

与传统的文档摘要相比,由于微博产品的特性及微博信息内容的特征,出现了一些和传统文档摘要以及新闻事件摘要不同的特点,根据这些特性生成更适应微博文档以及微博用户的摘要。

2.2.1 热点事件相关微博信息内容的特点

(1)微博信息属于短文本,而且和其他短文本相比长度更短。有些微博从句子层面看口语化比较严重、规范性差,但是词语的表达能力较强,而且一条微博包含的句子一般不会超过 4 个,所以不适合对单条微博文本做抽取式摘要工作。

(2)微博信息往往聚焦于事件的一个方面。微博信息即时性强、交互性强,而且由于文字长度限制,微博内容往往聚焦于事件的某个时间和某个方面的问题。不同于长篇的新闻报道,要求覆盖事件的方方面面。所以,通过单条微博难以概括事件的情况。

(3)对事件情况报道质量高的微博文本中事件要素比较完整,尤其是时间要素。不管是媒体、政府机构等权威账号还是其他用户的高质量微博,由于字数限制和传播需求,往往包含完整的事件要素,突出时间短语。并且事件要素完整,时间短语突出也能提高微博摘要的可读性。

2.2.2 微博热点事件摘要定义

目前对于微博事件摘要的研究比较少,仅看到 B. Sharifi、M. A. Hutton 和 J. Kalita 参考 Twitter 中的 WhatTheTrend 系统的热点事件摘要功能,将微博事件摘要定义为在某事件的所有微博集合中抽取与某事件最紧密相关的微博[10]。与该研究面对的情况一样,微博文本太短,所以本文不做单条微博的摘要抽取,选取整条微博作为摘要的一部分。然而,B. Sharifi 等研究没有考虑到热点事件中单条微博难以概括事件的特点,所以本文选取多条微博的集合作为摘要文本,同时按时间段划分微博集合。因此,本文将微博热点事件的摘要定义为在某时间段内所有微博中选取最能概括该事件情况的微博集合。

2.2.3 微博用户参与热点事件的行为特点

(1)微博对事件报道的质量和发布者身份相关。事件的新闻报道都是由记者撰写、新闻机构发布。新闻质量虽然与发布者水平相关,但是对事件情况的报道质量差距不大。然而微博由于没有审核过程,内容质量参次不齐。但是,由于微博用户拥有粉丝,或者本

身是政府机构、媒体、集团、网站、名人的社交账号，发布信息时需要承担社会责任或者满足粉丝期望，他们发布的微博质量很高，对事件某个方面的描述十分准确，是作为摘要的理想微博；同时有些事件的亲历者或见证人由于其微博的信息及时，质量高引起大量转发，也是选做摘要的理想微博。

（2）网民查看微博信息、参与微博事件讨论的频率按天呈周期性。根据《2015 年中国社交应用用户行为研究报告》[13] 显示，47.5% 的用户每天会看微博。用户对自己关注的热点事件一般会关注当天有哪些信息，昨天有哪些信息，因此，可以考虑按照每天组织微博摘要集，符合用户的使用习惯。

（3）网民对微博事件相关报道关注点不局限于一个方面，且在事件发展的时间内持续关注。新闻对事件的报道是聚合式的，但是微博对事件的报道往往是一个个方面，用户关注点不同，而且往往会跟进关注。如 2 月 6 日台湾高雄地震，当天微博的报道以及用户关注点包括地震地点、震源、伤亡情况等多方面，2 月 7 号会跟进关注救援情况，伤亡人数数据更新等。在浏览微博时会搜集各方面情况，由于微博数据量大，用户的信息需求往往难以满足。

综上所述，针对指定热点事件所有相关微博，按照每天划分为多个微博集合，再结合本文提出的复合模型抽取的摘要关键词将每天的微博集合中新闻价值高的微博筛选出来，形成摘要微博集，由于时间要素是每条微博中的重要组成成分，对每个摘要集中的每条微博，提取时间要素并利用本文提出的算法进行标准化，然后对摘要集中包含时间表达的微博，按照其提取的时间表达排序输出，对摘要集中不包含时间表达的置于最后。

2.3 LDA 模型介绍

LDA 模型基本原理是一个三层的贝叶斯模型，能够对文本中隐含的主题建模，与传统相似度计算等方法对比，LDA 可以无监督地从海量文本数据中自动生成语义主题。LDA 模型认为文档是主题的混合，将高维度的文本语料集合映射到低维度的潜在语义空间，认为主题是词空间上的分布，从而获得文本间的关系，描述的是一篇文档的产生过程。模型表示见图 2。

在 LDA 的生成过程中，对应的观测及隐藏变量的联合分布计算如公式(1)：

$$p(\beta_{1:K},\ \theta_{1:D},\ z_{1:D},\ w_{1:D}) = \prod_{i=1}^{K} p(\beta) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:k},\ z_{d,n})\right) \quad \text{公式(1)}$$

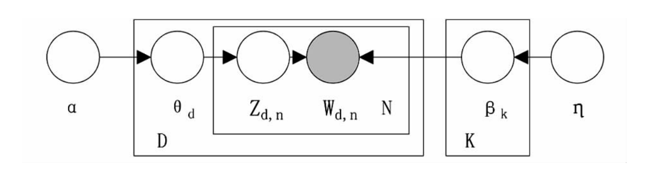


图 2　LDA 图模型

公式(1)中：β 表示主题，θ 表示主题的概率，z 表示特定文档或词语的主题，w 为词语。$\beta_{1:K}$为全体主题集合，其中 β_K 是第 k 个主题的词的分布（见图 2）。第 d 个文档中该主题所占的比例为 θ_d，其中 θ_d，k 表示第 k 个主题在第 d 个文档中的比例（见图 2）。第 d 个文档的主题全体为 Z_d，其中 $z_{d,n}$是第 d 个文档中第 n 个词的主题（见图 2 灰色圆圈）。第 d 个文档中所有词记为 w_d，其中 $w_{d,n}$是第 d 个文档中第 n 个词，每个词都是固定的词汇表中的元素。$p(\beta)$ 表示从主题集合中选取了一个特定主题，$p(\theta_d)$ 表示该主题在特定文档中的概率，$p(z_{d,n} | \theta_d$ 是该主题确定时该文档第 n 个词的主题，$p(w_{d,n} | \beta_{1:k},\ z_{d,n}$是该文档第 n 个词的主题与该词的联合分布。连乘计算随机变量的依赖性。

后验分布的计算见公式(2)：

$$p(\beta_{1:K},\ \theta_{1:D},\ z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K},\ \theta_{1:D},\ z_{1:D},\ w_{1:D})}{p(w_{1:D})} \quad \text{公式(2)}$$

在实际操作中，对于分子，给定的语料下很容易统计出来。分母计算量随着文本量的增大无法直接计算，如语料的综合词库超过百万，含有 n 个词语，每个词语计算 m 种观测组合，然后要累加得到先验概率，计算十分巨大。因此需要一种近似的求解方法。

常见的后验分布方法有 Expectation propagation，拉普拉斯近似以及吉布斯抽样等方法。本文采用的是吉布斯抽样方法来估计当前特征词和主题的后验分布。吉布斯抽样算法的流程是：第一步对语料中所有单词采样获取初始主题，第二步每当新观察到词语，计算当前主题，然后不断重复第二步直到所有主题的分布达到收敛。

最终获得到文档中多个主题下各特征词的概率[14]。

LDA 模型在摘要抽取上也有很多成功的应用。R. Arora 和 B. Ravindran 利用 LDA 模型计算每个主题下单词权重，以此获取每个句子的词语权重向量，使用奇异值分解算法获取最能表示主题含义的句子作为文档摘要，LDA 和 SVD 的混合模型很好地降低了摘要中的重复冗余部分[15]。Y. Petinot、K. Mckeown 和 K.

AThadani 提出 hLLDA 模型,建立每个标签与主题的对应关系,然后通过类别分层提取摘要[16]。

抽取式的摘要生成主要是从语料中抽取可以概括语料核心内容的总结性的句子,包括单文档处理和多文档处理的摘要抽取。LDA 模型可以建立多文档之间的语义联系,在应对语料信息稀疏,剥离冗余成分,提高摘要抽取的准确性上效果十分明显。

2.4 MaxEnt-MI 模型

在抽取微博摘要关键词的工作中,一般从词语间内部的紧密性和外部边界性来考察关键词之间的关系。内部紧密程度越高,说明该词组的完整性越好,即该词组内部词串联系紧密。外部边界可以衡量该词组表达整体的独立性,指数越高,该多字词表达的语义功能越强。如"发生地震""危机公关"。常见的内部方法有:t-score(在整体标准差不明的情况下,通过样本标准差来估测置信区间的 T 分布的坐标值)、互信息、对数可能性值等方法。常见的外部方法有:左右熵的方法等。

基于内部紧密性度量选择词语互信息模型,外部边界度量选择最大熵模型,组合两个模型提出衡量关键词间联系的的 MaxEnt-MI 模型,用于衡量关键词之间联系程度,该联系程度(定义为 MEMI)越高说明词语语义功能强而且词组完整性高。利用该模型处理语料的流程如图 3 所示:

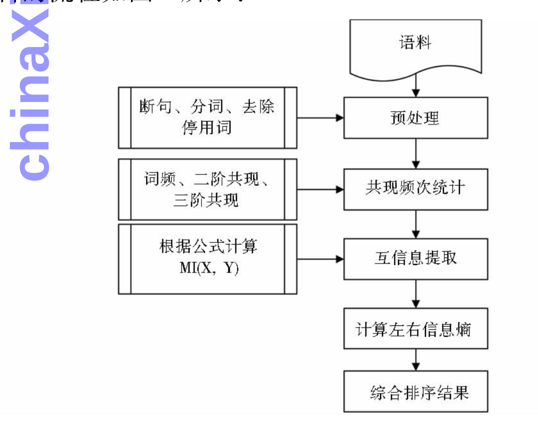


图 3 MaxEnt-MI 模型处理流程

2.4.1 词语互信息计算　互信息体现的是两个词之间相互依赖程度[17]。根据孙茂松和罗盛芬等学者在中文抽词任务中各种统计量的效果研究显示,互信息的抽词效果最好,而且多种方法之间的互补性强,在此研究基础上,选择使用互信息的统计量来判断词串之间的内部组合的相关程度。计算方法见公式(3):

$$\mathrm{MI}(X,Y)=\log_2\frac{P(X,Y)}{P(X)P(Y)} \qquad 公式(3)$$

其中,P(X)表示词 X 出现的概率,P(X,Y)表示词 X 和词 Y 在所有二阶词串中出现的概率,如"台湾→地震"语料中出现 314 次,所有的二阶词串一共有 2 000 个,那么 P(X,Y) = 314/2000。根据定义,互信息越高意味着两个词的内在结合的紧密程度越高,反之,两个词之间可能存在短语边界。

2.4.2 左右信息熵计算　一般来说,熵是用来衡量随机变量的不确定性[18]。即假设随机变量 X 可以取有限个随机变量,且 X 取这些变量的值都可以计算,其概率表示为 P(Xi),那么 X 的熵可以定义为 H(X),计算公式如下:

$$\mathrm{H}(X)=-\sum(x_i\in X)P(x_i)\cdot\log_2 P(x_i) \qquad 公式(4)$$

本文引用信息熵的定义来衡量多字词表达短语中左右边界的熵,以此来度量多字词短语的外部边界性。公式如下:

$$\mathrm{F}_L(\mathrm{W})=-\sum_{\forall a\in A}p(aW|W)\cdot\log_2 P(aW|W)$$
$$\mathrm{F}_R(\mathrm{W})=-\sum_{\forall b\in B}p(Wb|W)\cdot\log_2 P(Wb|W)$$

公式(5)

其中,F_L(W)和 F_R(W)表示目标词串组合的左熵和右熵,W 表示所有词串,A 表示目标词串左边出现过的词的集合,a 表示其中的某个词。B 表示目标词串右边出现过的词的集合,b 表示其中的某个词语。P(aW|W)表示词 a 出现在目标词串左边的概率。P(Wb|W)表示词 b 出现在目标词串右边的概率。

2.4.3 计算词语间联系程度　笔者将两个关键词能够组成语义功能性强度且内在联系程度高的短语的概率(MEMI)定义为以上 3 个统计量加权和,公式如下:

$$\mathrm{S(W)}=\alpha\cdot\mathrm{MI(W)}+\beta\cdot F_L+\gamma\cdot F_R \qquad 公式(6)$$

这一部分主要是根据已经获取的互信息值、左右信息熵对关键词组进行排序,权重的选择上主要是根据测试语料计算结果、人工评估效果,然后微调权重获取预估值,经过反复调试,参数 α 取 1,β 取 0.5,γ 取 0.6 时效果较好。经过测试,按照 MEMI 值排列得到的结果显示,每个词语只和某些词语有较强的联系,这些词往往是这个词语的固定搭配、动宾搭配、修饰关系等,语义功能性强。

2.5 LDA 模型和 MaxEnt－MI 模型的组合和改进

LDA 模型在处理微博语料时,可以很好地解决文本稀疏性的问题,而且无监督的方法使得在处理微博热点事件的时候,对领域知识的依赖少,但是 LDA 模型存在着以下几个问题,需要通过模型改进或者模型的组合来解决。

（1）LDA 模型在处理语料的时候最基本假设是，主题和文档中的词的顺序是无关的，即词项之间是可以交换的，因此，通过 LDA 提取的主题下关键词是一个无序的组合。然而关键词的次序对于事件非常重要，而且词语的前后顺序也影响了摘要的可读性。

（2）LDA 得出的是主题下各个词语可能属于该主题的概率，并没有涉及到该主题下词语的重要性。

（3）LDA 模型话题分布向高频词倾斜，导致很多能够代表话题的词语被高频词淹没，降低了模型的话题表达能力。

（4）LDA 模型需要设定主题的数量，主题数量设置不同，提取效果也不同[19]。

MaxEnt-MI 处理微博语料的时间复杂度低，效率高，在微博稀疏语料的条件下也很好地揭示了词语之间的联系，如前后顺序、固定搭配、动宾搭配、修饰关系等，具有语义功能性强、便于理解的效果。该模型只考虑了词语之间的特征，没有考虑词语与微博见的特征，忽略词语位置的关系以及语料中低频词的功能被放大，导致语料中表现能力较差的词语也被选作文本表示[20-21]。

但是 MaxEnt-MI 和 LDA 模型在某些方面很适合微博短文本摘要处理，并且两者有很好的互补性。MaxEnt-MI 解决了 LDA 模型词语的顺序问题以及提高了关键词的可读性；LDA 对语料进行主题提取后，生成的关键词聚类方便 MaxEnt-MI 进一步处理，在 LDA 模型获取的话题 - 关键词后，MaxEnt-MI 模型根据关键词的词性、权重等特征可以完成短语识别，结合事件要素等信息完成事件关键词筛选等工作，获取时间的摘要的关键词识别。

除了通过组合模型的方法之外，本文为了提高模型精度进行了一系列辅助性操作：①通过对所有语料进行分词与词性标注提取指定的实词列表，命名实体识别方法识别相关人名、地名、机构名要素，基于 TextRank 方法提取微博关键词，将以上 3 个词表去重综合形成本文的事件主题词表，根据词表过滤掉语义功能性比较差的词语（如停用词、连词等），这样可以提高表意性比较强的词语在 LDA 采用中的占比，从而获取更好的结果。②通过识别事件重要时间节点数确定 LDA 模型提取的主题数，使 LDA 提取的主题更加准确。③利用 MaxEnt – MI 最终输出的摘要关键词进行可读性优化。

3 实验过程

3.1 预处理

预处理主要是对采集的微博热点事件相关的微博进行分词、去除停用词、去除微博特殊符号（表情符号、话题符号、URL、@ 昵称的转发回复）、命名实体识别以及词性标注等。本文采用开源 ICTCLAS 分词系统完成分词、词性标注、命名实体识别，该系统支持用户自定义词典，分词速度快，准确率高[20]。

3.2 特征抽取

特征抽取部分有 3 个任务，获取事件相关的事件主题词表、获取事件重要时间节点和抽取事件要素。

3.2.1 获取事件主题词表　事件主题词由事件微博集中所有的实词组成。提取过程如下：①对所有语料进行分词与词性标注提取词性为名词、动词、形容词、副词的词语的实词列表。②使用 ICTCLAS 命名实体识别方法识别相关人名、地名、机构名要素。③利用 TextRank 方法提取每条微博关键词，在提取关键词的过程中，参照 R. Long 等的处理方法[8]，筛选处理后长度大于 10 的微博，对于微博句子，使用公式（7）确定提取关键词的个数。

$$N = \max\left\{\alpha, \frac{\frac{L}{\beta} + N}{2}\right\} \quad \text{公式(7)}$$

其中，L 表示句子长度，β 为压缩比例，N 为句子中名词实词的个数，α 为单句最少关键词个数。一般句子越长，名词实词越多，句子中包含关键词越多，经过试验，β 取 5，α 取 2 的时候效果最好。

将以上 3 个词表去重综合形成本文的事件主题词表，根据词表过滤掉语义功能性比较差的词语（如停用词、连词等）

3.2.2 重要时间节点表达式获取　热点事件发生后，微博上关于事件的报道中很多都会包含时间短语，这些时间短语对于事件摘要抽取和表示都尤为重要，是展示时间来龙去脉的重要信息。但是表达比较杂乱，需要统一处理和一定的逻辑推算：①采用 ICTCLAS 工具识别时间命名实体表达，如“2004 年 3 月 3 日”“昨日”“以前”，将结果分为两类，可以标准化和不能标准化的。②将可以标准化的时间结合自定义的规则进行标准化，然后按照时间推算。文本时间推算分两步：第一步，获取时间推算的两个要素（时间基准、时间关系短语），如果文中有时间短语关系（如“三天以后”“昨天下午”等），搜寻短语前面是否有时间基准（如“今天”，完整和不完整的时间表达）如果文中有时间关系

但是没有出现时间基准,以微博事件为基准。第二步,根据时间基准和时间关系短语推算时间,并将原文中的时间关系短语转换成标准时间。③统计从文中抽取的标准时间,获取重要时间节点集合。

以“高雄地震”事件全部微博(2016 年 2 月 6 日到 2016 年 2 月 15 日)作为数据集,时间短语统计样例及标准化如表 1 和表 2 所示:

表 1　时间短语抽取样例

任务	结果(部分)	
时间短语抽取	【原文】日本政府今天说,向中国台湾提供 100 万美元赈灾援助。日本始终没忘 5 年前的 311 大地震时中国台湾给日本的帮助 【基准时间】:今天 → 2016 年 2 月 8 日 【推算时间】:5 年前 → 2011 年	
时间短语统计	2016 年 2 月 6 日 3 时 57 分 1033	2016 年 2 月 6 日 10 时 40 分 95
	2016 年 2 月 6 日 1004	2016 年 2 月 9 日 4 时 93
	2016 年 2 月 7 日 166	2016 年 2 月 8 日 89
	2016 年 2 月 6 日 6 时 107	2016 年 2 月 7 日 12 时 30 分 85
	2016 年 2 月 9 日 8 时 47 分 101	2016 年 2 月 5 日 75

表 2　时间短语抽取部分规则

<table>
<tr><th>类型</th><th>匹配规则(部分)</th><th>举例</th></tr>
<tr><td>完整时间表达</td><td>yyyy-MM-dd-HH-mm-ss</td><td>2016-04-20-10-56-05</td></tr>
<tr><td>不完整表达</td><td>[0－9]? [0－9]{3}(? ＝年)</td><td>987 年,2010 年</td></tr>
<tr><td></td><td>((10)|(11)|(12)|([1－9]))(? ＝月)</td><td>2 月</td></tr>
<tr><td></td><td>((? <! \\d))([0－3][0－9]|[1－9])(? ＝(日|号))</td><td>1 号,12 日</td></tr>
<tr><td></td><td>(? <! (周|星期))([0－2]? [0－9])(? ＝(点|时))</td><td>9 点</td></tr>
<tr><td>模糊表达</td><td>\\d＋(? ＝年[以之]? 后)</td><td>一年后等</td></tr>
</table>

3.2.3　事件要素识别　事件要素是热点事件的重要组成成分,本文按照新闻学领域对于事件描述的 5W1H 抽取事件要素[22],即 where(地点)、when(时间)、who 和 whom(参与者)、what(具体动作)、how(结果),对于地点、人名、机构等命名实体可以直接使用 ICTCLAS 完成标注[20]。时间要素按照前面提出的时间短语识别的方法完成,what 也就是具体的动作通过句法分析和预定的规则模板(NP1＋V＋NP2、NP＋V、V＋NP、V＋Va 等)抽取主谓宾三元组,how 要素通过情感词典提取语句中情感倾向强的词语。

根据以上的方法,可以获取单个微博中事件要素的集合{T,L,NT,P,V,S},T 代表时间要素集合,L 代表地点等要素的集合,NT 代表机构地点等要素集合,P 代表人名职称等词语集合,V 代表动作词语集合,S 代表情感倾向词集合。处理文档为某时间段或者全部文档的时候,只需要把每个微博中的各要素整合即可。

获取时间段上的事件要素集合以及整体语料上的事件要素集合后,可以计算该时段上各要素的权重。权值量化公式如下:

$$W_F(f_i)=\frac{tf(f_i+T)}{sum(T)} \qquad \text{公式(8)}$$

其中,$W_F(f_i)$表示时间段内,事件要素 F 中词语 f_i 在要素 F 所在集合中的权重,$tf(f_i+T)$表示词语 f_i 在时间段 T 内出现的次数,$sum(T)$是 T 事件内所有事件要素 F 中的要素总个数。

3.3　组合模型获取事件摘要关键词

3.3.1　提取话题关键词　利用上一步提取的事件主题词表筛选微博语料,选取处理后长度大于 10 的语料作为训练 LDA 模型,并利用吉布斯方法采样[12],过程如下:①对文档集中每个文档的每个词随机赋予一个主题。②扫描文档集对每个词使用吉布斯采样公式重新采样主题,并在文档集中更新。③重复以上的重新采样过程直到收敛。④获取主题词汇概率矩阵。

3.3.2　关键词赋权重　对于抽取出来的主题关键词表,需要对其进行关键词权重计算。权重需要考虑关键词的两个重要属性:① 关键词对于事件的描述能力一般和词性以及词语充当事件要素的什么角色相关。如“地震”就是表现力比较强的词语。②该关键词在选定的时间段内出现频率比较高,但是在其他时间段内频率并不高,即衡量词语在时间跨度上的波动性,这里参考标准差的方法衡量。综合两个要素,通过以下公式计算关键词的权重 W_i。

$$W_i=\alpha*\sqrt{\frac{1}{T/\Delta T}*\sum_{k=1}^{T/\Delta T}\left(\left(F_k-\frac{F_i}{T/\Delta T}\right)^2\right)} \qquad \text{公式(9)}$$

其中,T 指热点事件微博的时间跨度,ΔT 是选取的时间间隔,F_k 是对应各时间段内该词语的词频,F_i 是词语的总词频,α 是该词语对应的事件要素。通过对数据预处理,观察各参数取值情况下关键词输出效果,最终确定地点取 0.75,人名取 1,机构取 0.5,时间取 0.75,其他词语取 0.2。

以“高雄地震”事件为例,选取 2016 年 2 月 6 日数据,LDA 主题关键词结果样例见表 3。

3.4　利用 MaxEnt－MI 模型合并关键词,生成关键词摘要

根据获取的关键词及其权重,选取 W_i ＞I 的关键词,根据 MaxEnt－MI 提取的词语关系词对(以台湾高雄地震时间为例,部分结果见表 4),将关键词分为两个集合,事件对象集 Subj{$Pair_1$,,⋯,,}

表 3　主题关键词结果样例(部分)

任务	结果				
提取主题关键词	topic 1	topic 2	topic 3	topic 4	topic 5
	台湾 = 0.048	地震 = 0.141	倒塌 = 0.065	影响 = 0.038	台湾 = 0.054
	高雄 = 0.031	台湾 = 0.098	地震 = 0.063	晚点 = 0.029	大陆 = 0.026
	地震 = 0.030	深度 = 0.028	大楼 = 0.036	厦门 = 0.028	提供 = 0.026
	报告 = 0.025	震源 = 0.025	台湾 = 0.034	列车 = 0.028	帮助 = 0.020
	受灾 = 0.023	高雄市 = 0.024	台南市 = 0.034	福州 = 0.024	灾情 = 0.019
	目前 = 0.018	测定 = 0.020	传出 = 0.030	铁路部门 = 0.017	需要 = 0.016
	发生 = 0.017	台网 = 0.019	建筑物 = 0.026	明显 = 0.014	海协会 = 0.016
	凌晨 = 0.016	震感 = 0.013	能量 = 0.026	震感 = 0.013	表示 = 0.016
	伤亡 = 0.01	附近 = 0.009	报道 = 0.024	杭深线 = 0.013	协助 = 0.011
	安全 = 0.01	高雄 = 0.009	呼救声 = 0.021	旅客 = 0.013	救灾款 = 0.010

事件行为集 Action{$Pair_1$,,…,,}集合构建的规则如下:①Subj－$Pair_i$ 选取关键词组合为:n＋n、a＋n(其中 n 按照细分词性标注选取 nt＊、nr、nn＊),并且加入命名实体名词。②Action－$Pair_i$ 选取关键词组合为:n＋v、v＋n、v＋v、v＋t、adv＋v、a＋v,并且加入动词实词。最终输出 Subj＋Action 词组组合的事件关键词摘要。

以“高雄地震事件”为例,选取 2016 年 2 月 6 日数据,MaxEnt－MI 模型抽取的关系词对结果以及事件关键词摘要样例如表 4 所示。

表 4　摘要关键词生成关键词(部分)

<table>
<tr><th>任务</th><th>结果</th></tr>
<tr><td>摘要关键词</td><td>Subj:台湾、台湾|高雄市、高雄、救援|人员、台南|地区、台南市|消防局、中国|地震局
Action:大楼|倒塌、人员|伤亡、等待|救援、地震|影响、祈福|台湾、马英九|痛斥</td></tr>
<tr><td>关系词对</td><td>台湾<-->地震　救援<-->人员　台湾<-->同胞　同胞<-->平安
台南<-->地震　高雄<-->台南　央视<-->新闻　同胞<-->祈福
台湾<-->高雄　高雄<-->地震　金龙<-->大楼　台湾<-->朋友
倒塌<-->大楼　台南<-->大楼　旅游<-->团队　大楼<-->救出
台湾<-->南部　大楼<-->倒塌　人员<-->伤亡　台湾<-->旅游</td></tr>
</table>

3.5　根据摘要关键词筛选微博作为事件轴摘要

在获取微博事件摘要的关键词后,针对每个时间段的微博,需要输出可以代表该时间内时间发展情况的句子集,这些句子需要反应该时间节点网络上针对该时间报道的主要内容。

本文抽取重要微博句子作为该时间节点上的摘要,与新闻摘要比,除了载体不一样外,诉求是一样的。因此,这里可以参考新闻价值的定义来制定句子的筛选标准[23]。新闻的价值是指新闻所含满足公众需求因素的总和,或称为社会价值的总和[24]。传播学中新闻价值五要素包括时效性、重要性、显著性、接近性、趣味性。按照新闻的时效性的事实原则,应该包含事件要素的情况(时间、人物、地点、动作)良好的新闻摘要句文本包含事件要素比较多;按照显著性,即考虑报道人的知名度,是否是关键传播节点(大 V,高转发微博等)也十分重要;按照接近性,微博包含事件关键词的情况,体现了微博和事件相关程度。因此在实现的时候需要综合考虑上述要素。本文结合新闻传播学新闻价值的定义,提出以下公式筛选摘要微博:

$$P(i)=\left(\frac{4E_T+2E_P+2E_L+2E_V}{10}\right)*\left(max\left(\frac{K_i}{K_t}\right)\right)*H_i$$

公式(10)

第一个大括号中是计算该微博包含事件要素的情况。E_T 是指是否包含标准化的时间短语,如果包含取 1,不包含取 0;E_P 指是否包含人名,E_L 指是否包含地名,E_V 是否包含动词实词。如果一个微博所有要素都包含,那么从新闻表达的角度看,它是新闻摘要的可能性很大。

第二个大括号内是通过对比该微博的关键词与本文组合模型提取的事件摘要关键词对比,K_i 表示该微博命中关键词个数,K_t 表示该主题内关键词个数。如果一个微博和主题的关键词重合越多,那么该新闻和该主题的相关性越大。

最后一个 H_i,表示该微博的社交价值,本文从微博新闻价值的角度考虑,认为 H_i 的大小和博主微博的类型、微博热度(评价、转发、点赞)、博主影响力相关。例如,名人在热点事件中发的微博可能在内容上不符合微博新闻价值的标准,但是名人的参与本身就是该热点事件的一部分,很多网民希望在了解事件的发展状况中看到该微博;还有,政府、官方媒体由于其本身的权威性,他们关于热点事件的报道可能比微博达人和普通用户的微博更能引起用户的兴趣。本文从微博事件摘要的角度提出了简化的微博社交价值衡量方法:①根据抓取用户账号的标签(企业、媒体、政府、名人、网站、团体、校园、达人、普通用户、其他)判断用户属于哪一类。如果用户属于高影响力类型(名人、政府、媒体、企业、网站),直接输出 Hi＝1。②如果用户

李纲, 徐伟, 王馨平. 基于事件要素的组合模型微博热点事件摘要提取[J]. 图书情报工作,2018,62(1):96－105.

属于中等影响力中等用户(团体、校园、达人)且转发数低于事件所有微博的一半,Hi＝0.6;高于所有微博的一半,Hi＝1。③如果用户属于普通用户和其他,转发数低于事件所有微博的一半,Hi＝0.4;高于所有微博的一半,Hi＝1。

以高雄地震事件为例,选取 2016 年 2 月 6 日数据,提取部分微博摘要如表 5 所示:

表 5　提取摘要效果(部分)

任务	结果
摘要关键词	Subj:台湾、台湾|高雄市、高雄、救援|人员、台南|地区、台南市|消防局、中国|地震局 Action:大楼|倒塌、人员|伤亡、等待|救援、地震|影响、祈福|台湾、马英九|痛斥
抽取微博摘要	事件时间:2016－2－6 03－57 【台湾高雄市发生 6.7 级地震震源深度 15 千米】中国地震台网正式测定:02 月 06 日 03 时 57 分在台湾高雄市(北纬 22.94 度,东经 120.54 度)发生 6.7 级地震,震源深度 15 千米。 时间:2016－2－6 03－57 #地震快讯#中国地震台网正式测定:02 月 06 日 03 时 57 分在台湾高雄市(北纬 22.94 度,东经 120.54 度)发生 6.7 级地震,震源深度 15 千米。(中国地震台网速报) 事件时间:2016－2－6 10－40 #高雄 6.7 级地震#【救出 225 人 5 人死亡】截至今天上午 10 时 40 分,台湾南部地震已造成 5 人死亡。救援人员已经救出民众 225 人,收容 74 人,送医 58 人。国台办表示,如需协助,大陆方面愿提供援助。愿平安!

4　实验结果分析

通过选取《2016 年度社会热点事件网络舆情报告》中具有代表性的事件"台湾高雄 6.7 级地震"和"和颐酒店女生遇袭",根据事件标题及扩展的查询词作为搜寻关键词,利用爬虫软件抓取微博,约 2 万条微博作为语料进行实验,微博采集样本及情况如表 6 所示:

表 6　微博采集样本

事件名称	采集关键词	采集时间	微博数目
台湾高雄 6.7 级地震	台湾高雄 6.7 级地震,高雄 6.7 级地震,高雄地震	2016 年 2 月 6 到 2016 年 2 月 15 日	7 737 条
和颐酒店女生遇袭	和颐酒店女生遇袭,和颐女生遇袭,女子入住北京酒店遇袭	2016 年 4 月 5 到 2016 年 4 月 16 日	11 166 条

由于摘要抽取的任务不同于一般的 NLP 任务,文摘评估的难点在于标准答案不唯一,对于一篇文章来说,表达同样内容的两句话都可能成为摘要的一部分,而且仅凭借表达方式的不同无法区分哪一句更适合。虽然有一些学者根据文本的特点提出了一些自动评价的手段但是由于其实现起来复杂,不在本文研究范围内。大多数评测都是通过人工内部评测的方式,人工生成标准摘要。本文也使用内部评测的方法来评测效果。由于人工评测对评测者的要求比较高,如果评测者文学素养不够,很可能会影响评测的结果,而且由于本文样本量比较大,人工评测需要一定的策略:

首先,本文选取的是从微博集合中抽取代表该时间段热点事件发展状况的微博句作为摘要,因此,人工评测仅需要测评人员从热点事件相关微博中选择自己认为可以作为摘要的微博即可,由于人工筛选的工作量大,所以本文仅以两个事件"高雄 6.7 级地震""和颐女生遇袭"作为评测数据。

其次,为了减少个人文学素养对评测的影响,利用百度百科关于热点事件的词条作为参考,这些词条是众多网友一起编写,准确率比较高。

人工评测完后,会获取到人工抽取的摘要集合作为测试集。本文通过常用的 F-measure 作为评测标准,如以下公式所示:

$$R = \frac{N}{N_p}$$

$$P = \frac{N}{N_r}$$

$$F - measure = \frac{2 * R * P}{R + P} \qquad 公式(11)$$

其中,N 是本文方法抽取摘要中符合人工抽取摘要的个数,N_p 是人工标注的摘要总数,N_r 本文方法抽取摘要的总数。

为了进行对比,本文将新闻摘要中 TextRank 算法自动摘要应用在微博语料中,某一时间段内的微博作为新闻文本,抽取微博摘要。如表 7 所示:

表 7　两种摘要方法对比

事件	召回率		准确率		F-Measure	
	本文方法	TextRank	本文方法	TextRank	本文方法	TextRank
台湾高雄 6.7 级地震	0.66	0.42	0.45	0.34	0.54	0.37
和颐酒店女生遇袭	0.62	0.37	0.47	0.32	0.53	0.34

从实验结果看,本文方法 F 值相对 TextRank 方法提高了 8%－13%,证明本文的方法提高了面向微博热点事件时间轴摘要的质量。而且从测试人员的反馈来看,关键词加上微博语句的摘要,在时间轴上很好地展现了事件的来龙去脉,在帮助用户了解热点事件以及事件监测等方面具有重要意义。

研究不足之处是,本文的抽取微博摘要的召回率很高,但是准确率提高不多,有一些热度并不高但是摘要价值很大的微博没有能够抽取出来,这一类微博往往不是热点微博,但是反映了事件的重要进展,如"民间救援组织公羊队一些来自浙江的成员抵达台南维冠大厦,并获准进入倒塌大楼区域参与救援"等。另外,本文方法在关键词选取等关键流程都有很多优化空间,需要进一步研究。

综上所述,笔者在基于传统的自动摘要和新闻摘要的研究基础上,结合新闻传播的特点,利用时间轴作为线索组织,以关键词加关键微博的摘要形式实现了对热点事件的时间轴摘要提取,实验结果表明本文方法相对 TextRank 方法召回率有显著提高,在实现过程中提出通过组合 LDA 模型和互信息最大熵模型来提高摘要关键词准确性和可读性等创新性研究方法,最终得到了满足用户信息需求的时间事件轴摘要。

参考文献:

[1] GOLDSTEIN J, KANTROWITZ M, MITTAL V, et al. Summarizing text documents: sentence selection and evaluation metrics [C]//Proceedings of the 22nd annual international ACM SIGIR conference research and development in information retrieval. New York: ACM, 1999:121 – 128.

[2] CANHASI E, KONONENKO I. Multi-document summarization via Archetypal Analysis of the content-graph joint model[J]. Knowledge and information systems, 2014, 41(3):821 – 842.

[3] CAI X, LI W. Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization[J]. IEEE transactions on audio speech & language processing, 2012, 20(5):1597 – 1607.

[4] 王红玲, 张明慧, 周国栋. 主题信息的中文多文档自动文摘系统[J]. 计算机工程与应用, 2012, 48(25):132 – 136.

[5] LUO Y, XIONG S. A combination scheme for distributed multi-document summarization[J]. Journal of intelligence, 2013, 64(1):94 – 102.

[6] INOUYE D. Multiple post microblog summarization[J]. Reu research final report, 2010(1):34 – 40.

[7] SWAN R, ALLAN J. Automatic generation of overview timelines [C]// International ACM SIGIR conference on research and development in information retrieval. Athens: DBLP, 2000:49 – 56.

[8] LONG R, WANG H, CHEN Y, et al. Towards effective event detection, tracking and summarization on microblog data[C]// International conference on web-age information management. Berlin: Springer-verlag, 2011:652 – 663.

[9] WAN X. TimedTextRank: adding the temporal dimension to multi-document summarization[C]// SIGIR 2007: proceedings of the, international ACM SIGIR conference on research and development in information retrieval. Amsterdam: DBLP, 2007:867 – 868.

[10] SHARIFI B, HUTTON M A, KALITA J. Summarizing microblogs automatically[C]// Humanlanguage technologies: the 2010 conference of the North American chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010:685 – 688.

[11] GAGLIO S, LO RE G, MORANA M. Real-time detection of twitter social events from the user's perspective[C]// IEEE international conference on communications. London :IEEE, 2015:1207 – 1212.

[12] WANG Y. Distributed Gibbs Sampling of Latent Topic Models: The Gritty Details[EB/OL]. [2017 – 04 – 10]. http://www.52ml.net/wp-content/uploads/2014/04/LDA-wangyi.pdf.

[13] CNNIC. 2015 年中国社交应用用户行为研究报告[R/OL]. [2016 – 02 – 11]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/sqbg/201604/P020160722551429454480.pdf,26.

[14] PORTEOUS I, NEWMAN D, IHLER A, et al. Fast collapsed Gibbs sampling for latent dirichlet allocation[C]// ACM SIGKDD international conference on knowledge discovery and data mining. Las Vegas: DBLP, 2008:569 – 577.

[15] ARORA R, RAVINDRAN B. Latent dirichlet allocation and singular value decomposition based multi-document summarization[C]// IEEE international conference on data mining. Pisa: DBLP, 2008:713 – 718.

[16] PETINOT Y, MCKEOWN K, THADANI K. Ahierarchical model of web summaries[C]// The meeting of the Association for Computational Linguistics: human language technologies, proceedings of the conference. Oregon: DBLP, 2012:670 – 675.

[17] 范小丽, 刘晓霞. 文本分类中互信息特征选择方法的研究[J]. 计算机工程与应用, 2010, 46(34):123 – 125.

[18] SHANNON C E, WEAVER W. Themathematic theory of communication[J]. Physics today, 1962:97 – 117.

[19] 张小平, 周雪忠, 黄厚宽,等. 一种改进的 LDA 主题模型[J]. 北京交通大学学报, 2010, 34(2):111 – 114.

[20] 张华平. NLPIR 汉语分词系统 [EB/OL]. [2014 – 01 – 15]. http://ictclas.nlpir.org.

[21] ZHU H D, ZHAO X H, ZHONG Y. Feature selection method combined optimized document frequency with improved RBF network [C]// Advanced data mining and applications, international conference, Adma 2009. Beijing: DBLP, 2009:796 – 803.

[22] 何玲, 胡小强, 袁玖根. 麦克卢汉媒体观下微媒体的 5W 分析[J]. 传媒, 2013(12):55 – 57.

[23] 杨保军. 论新闻价值关系的构成[J]. 国际新闻界, 2002(2):55 – 60.

[24] 郝雨. 回归本义的"新闻价值"研究[J]. 上海大学学报社会科学版, 2006, 13(6):69 – 74.

李纲，徐伟，王馨平.基于事件要素的组合模型微博热点事件摘要提取[J].图书情报工作,2018,62(1):96－105.

作者贡献说明：

徐伟：负责模型设计、数据分析、论文主体内容撰写与论文修改；

李纲：负责论文研究思路与论文修改指导；

王馨平：负责论文校对，数据分析。

Hot Event Summary on Micro-blog Generated by Multi Model Based on Event Elements

Li Gang　Xu Wei　Wang Xinping

School of Information, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] In order to help the readers understand the contexts of the news event on micro-blog platform and improve readability and accuracy of micro-blog event summary, we propose a method for extracting the event summary organized by time axis based on event elements. [**Method/process**] Based on the characteristics of micro-blog text, we combine both advantages and disadvantages of the LDA and mutual information maximum entropy model (MaxEnt-MI) and extract event summary keywords, screening micro-blog with micro-blog communication value and theme relevance and generating event summary in the form of time-keywords-mircro-blog. [**Result/conclusion**] Comparing with the traditional TextRank method in the artificially labeled test set, we find the F value increased by 8% to 13%, and the internal tests show that the readability of the abstracts is significantly improved. The number of experimental texts and test sets and the richness of the event need to be further expanded, and more weighting strategies should be considered in order to improve the accuracy of the abstracts. The experimental results and the test results show that the proposed method is feasible and effective, which can meet the needs of the users for the hot event summary information, and improve the accuracy of the micro-blog abstract extraction.

Keywords: text mining　event summarization　latent dirichlet allocation　mutual information maximum entropy model